融合降噪自编码器与 BPSO 的特征组合方法及其中医应用

黄灿奕，杜建强†，聂　斌，曾青霞，朱志鹏，喻　芳

(江西中医药大学 计算机学院，南昌 330004)

摘　要：离散二进制粒子群算法（BPSO）在各种离散优化问题中有着诸多优势，但其很容易由于非线性的问题陷入局部最优解，无法得到最佳特征子集。而降噪自编码器可通过多层非线性网络进行映射与重构，对中医药数据有良好的处理效果。因此提出了一种融合降噪自编码器与 BPSO 的特征组合方法，该方法主要是利用降噪自编码器对特征进行非线性映射形成超完备基，然后在超完备基中通过 BPSO 进行搜索，从而得到最佳特征子集。分别采用临床糖尿病数据集和 UCI 数据集进行分析处理，实验结果表明，融合降噪自编码器与 BPSO 的特征组合方法对中医药临床实验数据有较好的适应性。

关键词：降噪自编码器；BPSO；非线性；中医药

中图分类号：TP　　doi: 10.3969/j.issn.1001-3695.2018.03.0207

Application of feature combination method of denoising autoencoder and BPSO in TCM

Huang Canyi, Du Jianqiang, Nie Bin, Zeng Qingxia, Zhu Zhipeng, Yu Fang

(*School of Computy Jiangxi University of Traditional Chinese Medicine*, *Nanchang* 330004, *China*)

Abstract: The discrete binary particle swarm optimization algorithm has many advantages in various discrete optimization problems, but it is very easy to fall into the local optimal solution due to the nonlinear problem, and the best feature subset cannot be obtained. The noise reduction self-encoder can be mapped and reconstructed through a multi-layer nonlinear network, which has a good effect on Chinese medicine data. This paper proposes a feature combination method of fusion noise reduction self-encoder and BPSO. This method mainly uses noise reduction self-encoder to perform nonlinear mapping of features to form super-complete basis, and then searches in super-complete basis through BPSO. To get the best feature subset. The clinical diabetes datasets and UCI datasets were used for analysis and processing. The experimental results showed that the combination of fused-noise self-encoder and BPSO features a good adaptability to clinical experimental data of TCM.

Key words: Denoising Autoencoder; BPSO; Nonlinear; TCM

0　引言

在中医药领域的临床实验数据中，大多呈现出多成分、多靶点以及非线性的特点[1]，且由于数据的复杂性，特征之间存在强相关性和冗余性。因此无法采用传统的统计分析方法来阐述数据内部的量效关系，所以亟需一种能够解决多变量与非线性问题的数据分析方法，为科研工作者提供技术支撑。

离散二进制粒子群算法(binary particle swarm optimization, BPSO)是由 Kennndy 和 Eberhart[2,3]共同提出的一种扩展粒子群优化方法，常常被被生物信息、背包问题和图形图像等领域广泛应用。但是 BPSO 在特征选择过程中容易陷入局部最优值[4]，导致该算法不能筛选出最佳特征子集。而降噪自编码器(denoising autoencoder, DA)是由 Vincent 提出的一种改良方法[5,6]，在自编码器的基础上对原始数据加入噪音，进行多层非线性网络学习，经过无监督的逐层贪心训练与系统性的参数优化，从而能够提取、编码出具有鲁棒性较好的特征。文献[7]提出了一种自适应离散粒子群算法，引入排斥过程克服早熟收敛问题，然而吸引与排斥的参数设置难以确定；文献[8]提出了一种基于 SVM-RFE-BPSO 算法的特征选择方法，利用 SVM-RFE 快速去掉部分无关特征，然后以粒子群算法继续搜索最优特征子集，同样其参数的设置也是难以确定；文献[9]提出了一种带有高斯白噪声扰动的混合粒子群算法，引入自适应调整种群多样性的阈值，使其不易陷入局部最优值，然而新的约束条件也使得最优值极不稳定。

收稿日期：2018-03-27；**修回日期：**2018-05-08

作者简介：黄灿奕（1993-），男，硕士研究生，主要研究方向为医药数据挖掘及机器学习；杜建强（1968-），男（通信作者），教授，博士，主要研究方向为数据库与数据挖掘（jianqiang_du@qq.com）；聂斌（1972-）男，硕士，主要研究中医药信息及数据挖掘；曾青霞（1995-），女，硕士研究生，主要研究方向为医药数据挖掘及机器学习；朱志鹏（1990-），男，硕士研究生，主要研究方向为机器学习及医药数据挖掘；喻芳（1992-），女，硕士研究生，主要研究方向为医药数据挖掘及机器学习.

因此，本文将降噪自编码器与离散二进制粒子群算法组合优化，通过 DA 的三层网络结构对特征进行非线性映射，使得输入层与输出层的数据尽可能相似[10,11]，从而形成超完备基，并利用 BPSO 在超完备基上进行搜索，直至找到最优特征组合为止。该算法不仅可有效的去除冗余特征，同时还能防止局部最优值，从而建立一个适合中医药数据的分析模型。

1　融合 DA 的离散二进制粒子群算法（DA-BPSO）模型构建

降噪自编码器[12]是结合鲁棒性和腐化输入对自编码器进行修改的方法。其基本思想是先进行腐化处理，即将原始输入矩阵 $X=(x_1,x_2,x_3,...,x_n)$ 里的每个值 $x_i\,(i=1,2,3\cdots)$ 随即置为 0，使得部分数据的部分特征丢失，如 $\tilde{X}=(x_1,0,0,x_4...,x_n)$，然后将腐化后的数据 $\tilde{x}$ 通过映射方法：$f_\theta(x)=S(W_1x+B_1)$，映射成一个隐含层表达 Y；对隐含层数据 Y 利用重构方法：$g_{\theta'}(y)=S(W_2y+B_2)$，重构成输出层数据 Z；通过反复迭代训练，使得误差函数 $L_H(X,Z)$ 最小，从而尽可能保证 Z 近似于 X。

离散二进制粒子群算法采用二进制形式的编码，用一组 0 或 1 的二进制串表示粒子的位置(代表解空间的位置)，根据其适应度函数值对粒子进行优劣评价，粒子的速度和位置会依据适应度值进行调整，从而实现粒子在解空间中搜索最优解。但是 BPSO 在非线性的数据中进行寻优，很容易陷入局部最优值，所以传统的 BPSO 难以满足中医药临床实验数据的特点，而 DA 可以有效的去除特征冗余和解决非线性等问题，故本文将 DA 和 BPSO 方法结合，提出了一种融合降噪自编码器与 BPSO 的特征组合方法，这样既可以反映数据的本质特征，也可以防止陷入局部最优值。

DA-BPSO 方法先是在降噪自编码器中，结合概率分布对原始输入数据腐化处理，再将处理后的数据集进行非线性映射与重构，并对模型的网络参数通过最小化均方误差调优操作，使得模型的效果最好，从而形成一组超完备基；然后利用 BPSO 在这组超完备基中寻找最优特征组合。该方法不仅可以提高分类器的分类准确率和降低冗余特征对其的影响，还可以找到对因变量最重要的一组影响因子。DA-BPSO 构建过程如图 1 所示。

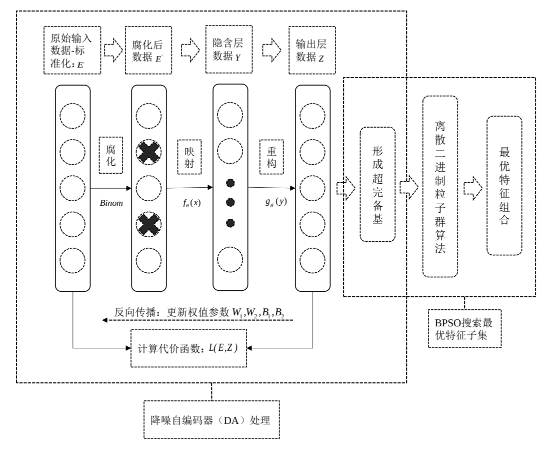


图 1 DA-BPSO 结构

具体步骤如下：

a)数据标准化预处理，原始数据 $X \Rightarrow$ 标准化数据 E。

b)腐化操作，先基于二项分布 binomial(0-1)产生一个腐化矩阵 $Binom$，再利用腐化矩阵对输入数据 E 进行腐化，即 $E'=E\cdot Binom$。

c)映射与重构，利用 Sigmoid 激活函数（式（1））将腐化后的数据 E' 映射成隐含层数据 Y，再将数据 Y 进行重构，得到输出层数据 Z。

$$sigmoid(\delta)=\frac{1}{1+e^{-\delta}} \tag{1}$$

d)计算代价函数与优化参数。利用输出层数据 Z 与输入数据 E 计算代价函数 $L(E,Z)$，并利用代价函数[14]的最小化均方误差进行参数优化，即通过反复迭代训练来更新网络中的参数，使得误差函数最小，从而得到效果较好的重构数据，形成超完备基。

最小化均方误差为

$$\theta,\theta' = \mathop{\arg\min}_{\theta,\theta'} \frac{1}{n}\sum_{i=1}^{n} L(e^{(i)},z^{(i)}) \tag{2}$$

$L(E,Z)$ 的损失函数为

$$L_H(e,z)=-\sum_{k=1}^{d}[e_k\log z_k+(1-e_k)\log(1-z_k)] \tag{3}$$

权重更新矩阵为

$$\begin{cases} W \leftarrow W-\varphi\cdot\dfrac{\partial L(e,y)}{\partial W} \\ B_1 \leftarrow b_y-\varphi\cdot\dfrac{\partial L(e,y)}{\partial B_1} \\ B_2 \leftarrow b_z-\varphi\cdot\dfrac{\partial L(e,y)}{\partial B_2} \end{cases} \tag{4}$$

e)初始化粒子的位置与速度。定义超完备基中的粒子数为 N，维度为 M，则粒子群 Q 可以表示为：$Q_i=(q_i^{(1)},q_i^{(2)},q_i^{(3)},...,q_i^{(M)}),i=1,...,N$（$Q_i$ 代表每个粒子的位置），粒子的初始值采用二项分布随机生成 0 或 1；每个粒子对应的速度 V_i 定义为：$V_i=(v_i^{(1)},v_i^{(2)},v_i^{(3)},...,v_i^{(M)})$，对粒子的初始速度随机初始在[0,1]之间。

f)计算适应度函数值。对每个粒子中被选中的特征，并基于超完备基抽中的数据，组成新的数据集，利用分类器进行分类，计算准确率 $accuracy(Q_i)$，从而得到每个粒子的适应度函数值：

$$f(Q_i)=-(A*accuracy(Q_i)+B*\frac{1}{n_features(Q_i)}) \tag{5}$$

其中：$n_features(Q_i)$ 是特征子集的数量（即每个粒子中 1 的个数），A 和 B 是权重参数（根据不同数据集进行调整，使得其在分类准确率和特征子集大小之间进行折中），取值范围在(0,1)之间。

g)记录每个粒子的历史最优适应度函数值 $f_{ih}(Q_i)$，以及最优适应度函数值时对应的个体最佳位置 $Q_{ih}=(q_i^{(1)},q_i^{(2)},q_i^{(3)},...,q_i^{(M)})$，直至记录全部粒子的最优适应度函数值 $f_{wh}(Q_i)$ 和对应的全体最佳位置 $Q_{wh}=(q_i^{(1)},q_i^{(2)},q_i^{(3)},...,q_i^{(M)})$；

h)对每个粒子的位置和速度进行更新，更新公式如下：

$$\begin{cases} V_i^{k+1} = wV_i^k + c_1\xi(Q_{ib}^k - Q_i^k) + c_2\eta(Q_{wb}^k - Q_i^k) \\ Q_i(t+1) = \begin{cases} 0, if(r_0 \ge Sig(V_i(t+1))) \\ 1, otherwise \end{cases} \\ Sig(v) = \dfrac{1}{1+exp(-v)} \end{cases} \tag{6}$$

其中：w 是保持原来速度的系数，称做惯性权重；c_1 和 c_2 是粒子跟踪自己历史最优的权重系数和粒子跟踪群体最优值的权值系数，c_1 和 c_2 是通常设置在[1,2]之间（经过多次实验将其设置为 2）；ξ 和 η 是[0,1]区间的均匀分布随机数，Q_i^k 为第 i 个粒子在第 k 步的时的取值，t 为迭代次数，r_0 是均匀分布的随机数，$sig(v)$ 是激活函数；

i)重复寻优过程，直至找到最优位置，该位置即为最优特征组合，算法终止。

2　实验结果及分析

2.1　实验数据说明

本文的实验数据主要来源于江西中医药大学重点实验室的临床糖尿病数据和 UCI 数据集上的 Wine Quality、CASP。其中 Wine Quality 数据中有 11 个特征，1600 个样本；CASP 有 9 个特征，45730 个样本；临床糖尿病实验数据共有特征数 16 个，样本数 284 个，其特征主要有：BMI（指偏胖、肥胖、正常和瘦）、入组 HDLc、入组胆固醇、入组甘油三酯（TG）、入组胰岛素（oh）、入组糖化血红蛋白、入组收缩压、入组空腹静脉血糖、黄连分组编号等。部分实验数据如表 1 所示。

表 1　临床糖尿病部分数据

BMI	HDLc	胆固醇	TG	0h	组号	…
24.4	1.52	4.89	1.69	114.33	A	…
25.1	1.38	5.02	3.14	56.2	A	…
30.8	1.09	5.35	9.91	53.79	A	…
…	…	…	…	…	…	…
30.2	1.19	5.54	1.51	81.32	B	…
27.1	1.12	5.08	1.53	43.95	B	…
24.6	0.88	6.35	20.11	152.8	B	…
…	…	…	…	…	…	…
23.2	1.13	6.25	3.78	392.52	C	…
21.9	1.6	5.02	0.97	74.96	C	…
…	…	…	…	…	…	…

2.2　实验过程和结果分析

为了验证 DA-BPSO 的有效性，采用原始数据的全部特征（下面称为原始特征）与 4 种策略进行比较，策略分别有传统的 BPSO 搜索的特征组合、混合粒子群算法(HDPSO)搜索的特征组合、自适应离散粒子群算法(SADPSO)搜索的特征组合和 DA-BPSO 搜索的特征组合。采用上述不同策略对三个数据集进行特征分析，得出的结果为：

a)临床糖尿病数据。BPSO 选择策略后特征数有 8 个，HDPSO 选择策略后特征数有 11 个，SADPSO 选择策略后特征数有 8 个，而 DA-BPSO 由于进行 DA 操作时，将特征映射成 30 维，映射后的特征数为 30 个，所以其选择后的特征数有 11 个，如图 2 所示。

b)数据集 Wine Quality。BPSO 选择策略后特征数有 4 个，HDPSO 选择策略后特征数有 7 个，SADPSO 选择策略后特征数有 5 个，同样的由于进行 DA 处理时将特征映射成 20 维，映射后的特征数为 20 个，所以 DA-BPSO 选择后的特征数有 12 个，如

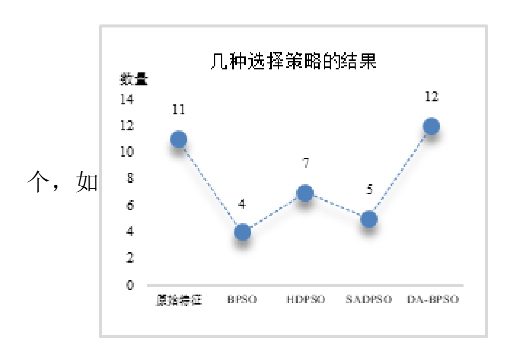


图 3 所示。

c)数据集 CASP。BPSO 选择策略后特征数有 2 个，HDPSO 选择策略后特征数有 5 个，SADPSO 选择策略后特征数有 4 个，同样的由于进行 DA 时将特征映射成 20 维，映射后的特征数为 20 个，所以 DA-BPSO 选择策略后的特征数有 7 个，如

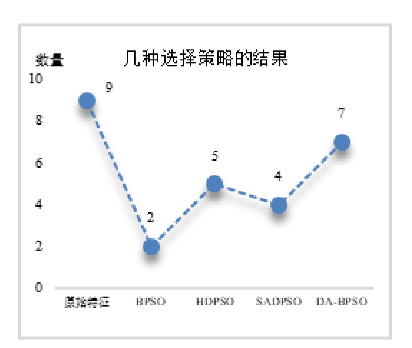


图 4 所示。

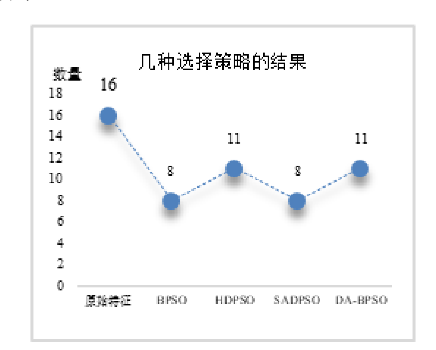


图 2　临床糖尿病数据的特征数量

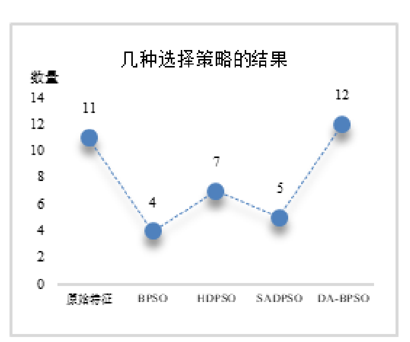


图 3 Wine Quality 的特征数量

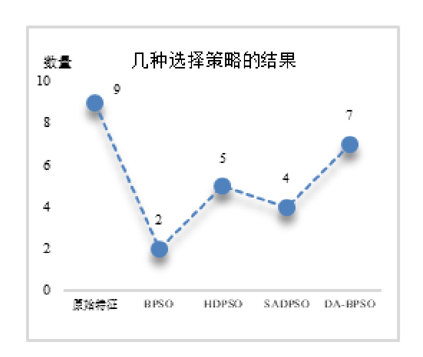


图 4 CASP 的特征数量

为了进一步分析改进算法的效果，本文选择以 SVM 作为分类器，分别以训练集（Train）和测试集（Test）的准确率作为比较，其中三个数据集都以 7:3 的比例随机划分成训练集和测试集，即 70%构建学习训练集，30%做测试集。由于在特征选择的时候，均是每次迭代 1000 次的最优结果，因此为了防止局部最优值的扰动，对每种策略各运行 10 次选择最优的特征组合。比较结果如表 2 所示。

表 2 原始特征与 4 种策略的实验结果比较

	原始特征		BPSO		HDPSO		SADPSO		DA-BPSO	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Wine Quality	0.5957	0.5267	0.5628	0.5506	0.5533	0.5628	0.5828	0.5742	0.6014	0.5933
临床糖尿病数据	0.3600	0.1056	0.3566	0.1227	0.3266	0.1702	0.3433	0.2304	0.3733	0.2978
CASP	0.7312	0.6415	0.7165	0.6548	0.6849	0.6654	0.7512	0.7054	0.7489	0.6900

根据上表可知，采用 SVM 作为分类器，原始特征的训练集准确率分别为 0.5957、0.3600、0.7312，测试集的准确率分别为 0.5267、0.1056、0.6415。为了更直观的显示实验结果，分别绘制了图 5 和 6，以体现训练集和测试集的准确率波动情况。由结果分析图可知，在训练集上原始特征的效果相对于 BPSO 与 HDPSO 都要好，然而测试集上就相对差很多，主要原因就是特征之间的冗余对测试集分类准确度的影响。

同时可以发现 SADPSO 在减少了特征数量的情况下，相比 HDPSO 在训练集和测试集的准确率都略有所提升。而 DA-BPSO 在特征数量较多时，训练集的准确率分别为 0.6014、0.3733、0.7489，测试集的准确率分别为 0.5933、0.2978、0.6900，相比 HDPSO、SADPSO 两种策略而言，虽然 DA-BPSO 映射后的特征数较多，但其准确率都表现出比较好的效果，说明改进的算法具有较强的鲁棒性。同时与原始特征的准确率相比较可得知，DA-BPSO 的效果均比其要好，特别是在临床糖尿病数据集上，呈上升趋势（如图 6 所示）。综上所述，改进的算法对具有非线性特点的中医药临床实验数据适应性良好，且能够防止陷入局部最优，从而得到最佳特征子集。

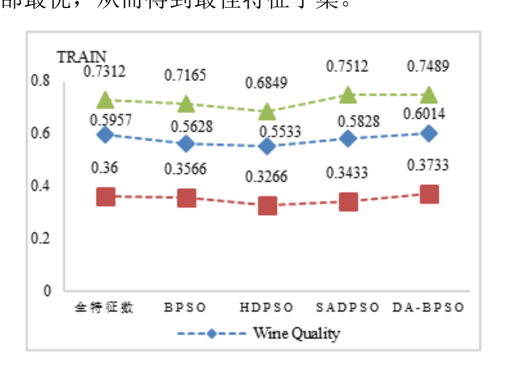


图 5 原始特征与 4 种策略的实验结果(Train)

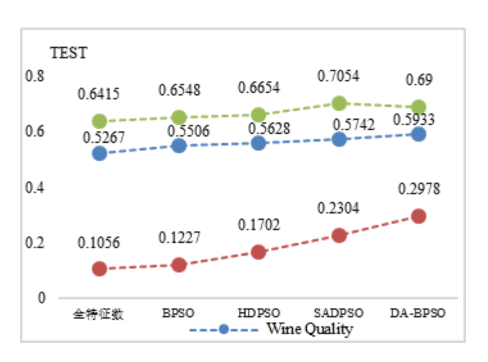


图 6 原始特征与 4 种策略的实验结果(Test)

3 结束语

本文针对在具有非线性特点的中医药临床数据中，传统的离散二进制粒子群算法容易陷入局部最优解，导致无法得到最佳特征子集的问题，提出了融合降噪自编码器与 BPSO 的特征组合方法，充分利用降噪自编码器在构建模型时加入噪音，同时进行非线性映射与重构获取超完备基等优点，并结合 BPSO 算法寻找最优特征组合，从而可有效的防止陷入局部最优解，增强模型的鲁棒性和泛化性。通过在临床糖尿病数据和 UCI 数据集的实验比较，证明该改进的算法明显提高了模型的分类精度和非线结构的表达，是一种适合于中医药领域的数据分析方法。但改进的算法也存在不足之处，其隐含层的个数会影响算法速度，从而导致需要更多的迭代次数才能达到最优效果。在接下来的工作中，将继续提升算法的搜索效率，同时在构建模型时如何保证相关参数的合理设置还可做进一步的研究。

参考文献：

[1] 张伯礼，王永炎. 组分配伍研制现代中药的理论与实践：方剂关键科学问题的基础研究 [M]. 沈阳：辽宁科学技术出版社, 2010. (Zhang Boli, Wang Yongyan. Research and development of modern Chinese medicine by multi-components compatibility: theory and practice: fundamental research on key scientific issues of prescriptions [M]. Shenyang: Liaoning Science

Technology Press, 2010.)

[2] 李浩君，刘中锋，王万良. 采用弧形映射函数的二进制粒子群优化算法 [J]. 小型微型计算机系统, 2017, 38 (12): 2637-2640. (Li Haojun, Liu Zhongfeng, Wang Wanliang. Binary particle swarm optimization algorithm with arcshaped transfer function [J]. Journal of Chinese Computer Systems, 2017, 38 (12): 2637-2640.)

[3] J. Kennedy, R. C. Eberhart. A discrete binary version of the particle swarm algorithm [J]. 1997, 5: 4104-4108 vol. 5.

[4] 毕凯，王晓丹，邢雅琼. 基于改进 BPSO 的聚类选择性集成 [J]. 系统工程与电子技术, 2016, 38 (3): 692-698. (Bi Kai, Wang Xiaodan, Xing Yaqiong. Cluster ensemble selection based on improved BPSO [J]. Journal of Systems Engineering and Electronics, 2016, 38 (3): 692-698.)

[5] Vincent P. A connection between score matching and denoising autoencoders. [J]. Neural Computation, 2011, 23 (7): 1661-74.

[6] Zhang Z, Jiang T, Li S, *et al.* Automated feature learning for nonlinear process monitoring: an approach using stacked denoising autoencoder and k-nearest neighbor rule [J]. Journal of Process Control, 2018, 64 (April 2018) .

[7] Dengli B U. MPRM minimization algorithm based on SADPSO [J]. Journal of Chongqing University of Posts & Telecommunications, 2016.

[8] 林俊，许露，刘龙. 基于 SVM-RFE-BPSO 算法的特征选择方法 [J]. 小型微型计算机系统, 2015, 36 (8): 1865-1868. (Lin Jun, Xu Lu, Liu Long. Feature selection method based on SVM-RFE and particle swarm optimization [J]. Journal of Chinese Computer Systems, 2015, 36 (8): 1865-1868.)

[9] 刘衍民. 一种求解约束优化问题的混合粒子群算法 [J]. 清华大学学报: 自然科学版, 2013 (2): 242-246. (Liu Yanmin. Hybrid particle swarm optimizer for constrained optimization problems [J]. Journal of Tsinghua University: Science and Technology, 2013 (2): 242-246.)

[10] Lore K G, Akintayo A, Sarkar S. LLNet: A deep autoencoder approach to natural low-light image enhancement [J]. Pattern Recognition, 2016, 61: 650-662.

[11] 王宪保，何文秀，王辛刚，等. 基于堆叠降噪自动编码器的胶囊缺陷检测方法 [J]. 计算机科学, 2016, 43 (2): 64-67. (Wang Xianbao, He Wenxiu, Wang Xingang, *et al.* Capsule defects detection based on stacked denoising autoencoders [J]. Computer Science, 2016, 43 (2): 64-67.)

[12] Li Peng, Ning C. Feature learning for music auto-tagging using denoising autoencoder [J]. Journal of East China University of Science & Technology, 2017, 43 (2): 241-247.

[13] 李阳辉，谢明，易阳. 基于降噪自动编码器及其改进模型的微博情感分析 [J]. 计算机应用研究, 2017, 34 (2): 373-377. (Li Yanghui, Xie Ming, Yi Yang. Sentiment analysis of micro-blogging based on DAE and its improved model [J]. Application Research of Computers, 2017, 34 (2): 373-377.)